

A Scalable Attention-Driven Architecture for Modeling Complex Disease Transmission Dynamics at Population Scale

K. Madhavi¹, Potharaju Gouthami¹, Pabbath Reddy Pavan Kumar Reddy¹, K. Sri Ram Venkata Mani Pavan¹, Mohammad Sammad¹

¹Department of Computer Science and Engineering, ¹Sree Dattha Institute of Engineering and Science, Nagarjuna Sagar Road, Sheriguda, Ibrahimpatnam, Rangareddy Dist, 501510, Telangana, India.

ABSTRACT

The rapid expansion of unstructured clinical text within healthcare systems has created a critical need for automated solutions capable of extracting meaningful insights for clinical decision support. Medical transcription data contains essential information regarding patient conditions, treatments, and associated domains, yet its unstructured nature makes accurate interpretation complex. The problem addressed in this study is the automatic classification of clinical narratives into appropriate medical specialties. In a manual system, healthcare professionals or administrative staff read and interpret each document individually to assign categories based on their expertise, which is time-consuming, inconsistent, and unsuitable for large-scale data processing. This manual approach suffers from limitations such as lack of scalability, high dependency on human judgment, susceptibility to errors, and absence of standardization, thereby emphasizing the need for an efficient automated system. To overcome these challenges, the proposed framework integrates natural language processing techniques with advanced machine learning methods. Initially, text data is pre-processed using tokenization, stop-word removal, and lemmatization to generate clean input. Contextual feature representations are then extracted using Efficiently Learning an Encoder that Classifies Token Replacements Accurately (ELECTRA), capturing semantic relationships within the text. To handle class imbalance, Synthetic Minority Over-sampling Technique (SMOTE) and K-Means Synthetic Minority Over-Sampling Technique (KMeansSMOTE) are applied. The resulting features are utilized to train multiple classifiers, including Adaptive Boosting Classifier (ABC), Extremely Randomized Trees Classifier (ETC), Random Forest Classifier (RFC), and Tree Alternating Optimization Classifier (TTC), enabling robust comparative analysis and improved predictive performance. The proposed system significantly enhances classification accuracy, reduces manual workload, and provides a scalable and reliable solution for intelligent healthcare text analytics.

Key words: Clinical Text Classification, Medical Transcription Analysis, Natural Language Processing (NLP), Ensemble Learning, Healthcare Text Analytics. Synthetic Minority Over-sampling Technique (SMOTE).

1. INTRODUCTION

The continuous emergence and re-emergence of infectious diseases, including H7N9, H5N1, Zika, and Ebola, represent a persistent and evolving threat to global public health due to their rapid transmission dynamics and, in many instances, high case fatality rates. A major challenge associated with these diseases is the limited availability of effective therapeutic interventions and vaccines, which significantly increases their potential to cause widespread health crises [1]. As a result, early detection of outbreaks and continuous monitoring of disease progression have become indispensable components of modern public health infrastructure. Accurately identifying anomalous patterns in disease occurrence, understanding transmission trends, and evaluating outbreak risks are critical for implementing timely containment and mitigation strategies. To address these requirements, public health authorities at national, state, and local levels have developed and deployed comprehensive disease surveillance systems aimed at improving situational awareness, preparedness, and response capabilities.

The growing interconnectedness of the world, driven by increased international travel and population mobility, has further intensified the complexity of disease spread across regions and borders. Pathogens can now disseminate rapidly from one geographic location to another, often before traditional detection mechanisms can respond effectively. In response, countries such as India have strengthened their surveillance efforts by systematically collecting and analyzing outbreak-related information from credible international organizations and health agencies [2]. However, the traditional approach of manually reviewing extensive volumes of reports, bulletins, and epidemiological updates on a daily basis is highly labor-intensive, time-consuming, and prone to subjective interpretation, resulting in inconsistencies and delayed decision-making processes. These limitations underscore the urgent need for automated, scalable, and reliable analytical solutions [3].

The widespread availability of digital information through online news platforms, health portals, and social media channels has introduced new opportunities for real-time disease monitoring and early warning systems. Contemporary web-based biosurveillance platforms leverage these heterogeneous data sources to detect emerging outbreaks, monitor disease progression, and assess regional risk levels more effectively [4]. The integration of Artificial Intelligence (AI) techniques further enhances these systems by enabling rapid processing of large-scale, unstructured textual data, extracting meaningful patterns, and identifying early signals of potential outbreaks. In this context, automated text classification serves as a foundational component, allowing the system to filter, categorize, and prioritize relevant information from vast data streams. This significantly improves the speed, accuracy, and scalability of outbreak detection, thereby supporting proactive public health interventions and reducing the overall impact of infectious disease outbreaks on global populations [5].

2. LITERATURE SURVEY

Gao, et al. [6] proposed a data- and knowledge-driven Named Entity Recognition framework that integrated Bidirectional Long Short-Term Memory with Conditional Random Field architecture along with multi-head self-attention mechanisms to capture both contextual and sequential dependencies in text. Their approach incorporated domain-specific knowledge bases to enhance entity recognition for components such as applications, vendors, and system attributes within cybersecurity datasets. By combining deep contextual embeddings with structured knowledge and attention-based feature refinement, the model effectively reduced ambiguity and improved precision in entity extraction tasks, particularly in domain-constrained environments. Meinert, et al. [7] conducted an extensive investigation into cybersecurity risks in healthcare systems, focusing on vulnerabilities associated with interconnected medical devices, electronic health record systems, and cloud-based infrastructures. Their study evaluated existing security protocols and identified limitations in traditional defense mechanisms when addressing advanced persistent threats and ransomware attacks. The authors emphasized the importance of integrating intelligent monitoring systems and adaptive security frameworks supported by artificial intelligence to improve threat detection and system resilience. Atal, et al. [8] evaluated a medical text classification system using a gold-standard external dataset and reported high sensitivity and specificity across multiple disease categories. Their framework successfully classified clinical trial records into a structured multi-class taxonomy aligned with the Global Burden of Disease classification. The study demonstrated the effectiveness of automated classification systems in handling large-scale clinical data and reducing reliance on manual annotation. Zhou, et al. [9] developed a hybrid Named Entity Recognition model that combined Bidirectional Encoder Representations from Transformers with Bidirectional Long Short-Term Memory and Conditional Random Field architectures. The transformer component captured deep contextual semantics, while the sequential layers modeled long-range dependencies and label transitions. This integrated architecture improved entity recognition performance in domain-specific datasets with complex linguistic structures.

Bui, et al. [10] introduced an Enrichment by Topic Modeling approach based on latent Dirichlet allocation to enhance semantic representation of short texts. The method incorporated probabilistic

topic distributions into textual features and adapted to variations in text length, enabling more robust representation of sparse and limited textual inputs commonly found in real-world datasets. Richter-Pechanski, et al. [11] investigated medical text classification as a fundamental task in medical natural language processing, focusing on categorizing short clinical texts into predefined classes such as disease stages, treatment conditions, and organ status. Their study demonstrated that accurate classification significantly influenced the performance of downstream applications, including adverse event detection and clinical decision support systems. Andreas Puder, et al. [12] proposed a self-optimizing framework for predicting cyber risks in healthcare systems through real-time algorithmic analysis. Their approach utilized adaptive models capable of continuously monitoring system behavior and identifying potential vulnerabilities and bottlenecks. The study emphasized the integration of artificial intelligence into healthcare infrastructures to enable proactive risk management, improve system reliability, and support secure digital healthcare operations.

Lena, et al. [13] investigated the application of artificial intelligence in requirements engineering within healthcare systems, focusing on improving system design and development processes. Their work analyzed limitations in existing methodologies and proposed domain-specific enhancements to address the complexity of healthcare applications. The study highlighted the role of intelligent techniques in improving requirement analysis, system efficiency, and overall development accuracy. Gromadzka, et al. [14] developed the Unified Wilson's Disease Rating Scale to evaluate a comprehensive range of clinical symptoms associated with Wilson's disease. Their study assessed the reliability and consistency of the scale using statistical measures, reporting high internal consistency and strong interrater agreement. The results demonstrated the effectiveness of the scale as a reliable tool for clinical assessment and patient monitoring. Fahn, et al. [15] analyzed clinical trial data involving patients with Parkinson's disease to determine the minimal clinically important change using the Unified Parkinson's Disease Rating Scale. Their study evaluated treatment effectiveness over a six-month period and established standardized thresholds for meaningful clinical improvement. The findings provided valuable insights for assessing disease progression and evaluating therapeutic outcomes.

3. PROPOSED SYSTEM

The proposed system architecture presents a unified and scalable pipeline for automated clinical text classification by integrating Natural Language Processing, transformer-based feature extraction, class balancing, and multiple ensemble Machine Learning models into a cohesive framework.

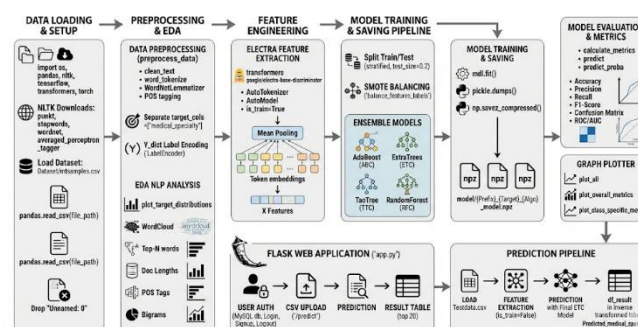


Fig. 4.1: Proposed system architecture

The process begins with structured data ingestion from MT Samples, followed by an extensive preprocessing stage where text is normalized through tokenization, stop word removal, and lemmatization to eliminate noise and ensure consistency. Target labels are encoded using label encoding, and Exploratory Data Analysis is performed to examine linguistic patterns such as word distributions, document lengths, part-of-speech tags, and n-grams, enabling better understanding of the dataset. The cleaned text is then transformed into dense contextual representations using Efficiently Learning an Encoder that Classifies Token Replacements Accurately, which captures deep semantic

and syntactic relationships; these embeddings are aggregated using mean pooling to produce fixed-length feature vectors suitable for machine learning models. To address class imbalance, SMOTE is applied, ensuring balanced representation of all medical specialties and improving generalization. The system incorporates multiple ensemble classifiers, including ABC, ETC, TTC, and RFC, each trained on the same feature space to enable robust comparative analysis and optimal model selection. The dataset is split using stratified sampling to preserve class distribution, and trained models are serialized for reproducibility and deployment. Performance evaluation is conducted using metrics such as accuracy, precision, recall, F1-score, confusion matrix, and multi-class ROC-AUC curves, providing a comprehensive assessment of classification effectiveness, as illustrated in Fig. 1, while visualization modules generate comparative performance graphs for deeper interpretability. The architecture is further extended with a Flask-based web application that supports real-time prediction by allowing users to upload new datasets, which undergo the same preprocessing and feature extraction steps before inference, ensuring consistency between training and deployment. The system achieves seamless integration of data processing, contextual embedding, ensemble learning, evaluation, and deployment, making it highly effective for real-world clinical decision support and automated medical specialty classification.

3.1 ELECTRA TRANSFORMER MODEL

The ELECTRA transformer model as depicted in Fig. 2 is an advanced pretraining paradigm designed for efficient and effective language representation learning. This architecture is composed primarily of two transformer models: a smaller generator and a larger discriminator, both based on stacks of transformer encoder layers utilizing multi-head self-attention and feed-forward networks. The generator's role is to predict plausible token replacements for a subset of input tokens, creating a corrupted input sequence. The discriminator then takes this corrupted sequence and predicts, for each token, whether it is original or replaced, training on a replaced token detection task rather than traditional masked language modelling.

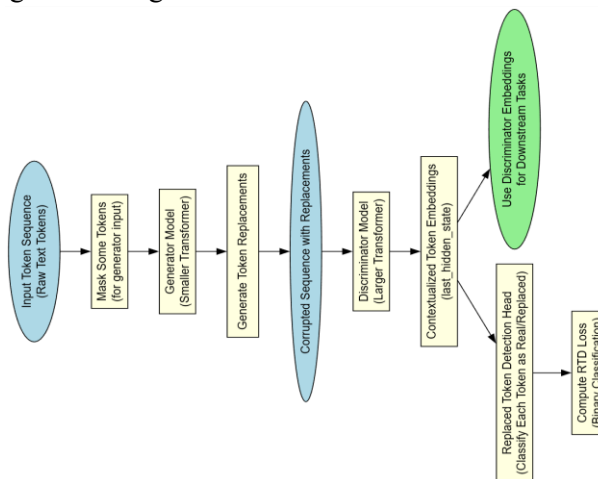


Fig. 2: Internal working operation of ELECTRA transformer model.

Each transformer encoder block within ELECTRA consists of an input embedding layer combined with positional encodings, followed by multiple layers of self-attention mechanisms that allow every token to attend to others in the sequence to capture contextual relationships. This is followed by feed-forward neural networks that apply non-linear transformations. Residual connections and layer normalization are employed throughout to stabilize training and enhance gradient flow. The discriminator produces rich contextual embeddings capturing deep semantic and syntactic information from the input text. Unlike models like BERT that mask tokens during training, ELECTRA's replaced token detection trains the model to discriminate real from fake tokens throughout the sequence, resulting in more efficient learning and better utilization of input data. In this work, the pretrained ELECTRA discriminator model

is used to generate deep contextual embeddings from the medical text, which are then pooled and used as numerical features for infectious disease classification.

4. DATASET DESCRIPTION

The dataset utilized in this research consists of clinical text data collected from publicly available medical case reports, clinical trial summaries, and disease-related documents. These records encapsulate real-world descriptions of patient conditions, diagnostic observations, and medical specializations, providing a rich source of unstructured text suitable for NLP-driven disease classification. Each record corresponds to a clinical document annotated with its relevant medical specialty, enabling supervised training of classification models. The dataset serves as an ideal benchmark for exploring transformer-based embeddings like ELECTRA to capture complex linguistic and semantic patterns within medical narratives.

- **sample_name:** A unique identifier assigned to each medical document or clinical text sample. It ensures record traceability and prevents duplication within the dataset.
- **description:** The main textual field containing unstructured medical narratives such as patient symptoms, diagnostic notes, treatments, or disease summaries. This column forms the core input for NLP preprocessing and feature extraction.
- **medical_keywords:** A curated list of domain-specific keywords extracted from the description, including disease names, anatomical terms, or medical procedures. These keywords aid in enhancing the semantic representation of the text.
- **source:** Specifies the origin of the clinical text (e.g., PubMed, WHO outbreak database, or clinical trial repository). This information helps track data provenance and assess linguistic variations across sources.
- **language:** Indicates the language of the document (e.g., English, Spanish, French). It enables multilingual processing and language-based analysis for model generalization.
- **report_date:** Represents the date on which the report or document was published or recorded. This temporal attribute supports chronological analysis of outbreak trends.

medical_specialty (Target): The target label denoting the specific medical domain associated with the text, such as Infectious Diseases, Cardiology, Neurology, or Respiratory Medicine. This column is used for supervised learning during disease classification.

4.1 RESULTS AND DISCUSSION

The results and discussion section presents the performance evaluation of the proposed medical text classification system developed using NLP, ELECTRA-based feature extraction, and multiple machine learning models. The system was trained and tested on a clinical text dataset, where advanced preprocessing and SMOTE-based balancing significantly improved data quality and class distribution. Among the implemented models, the Extra Trees Classifier demonstrated superior performance compared to AB, TT, and RF in terms of accuracy, precision, recall, and F1-score. The use of transformer-based embeddings enhanced semantic understanding of medical text, leading to more reliable predictions.

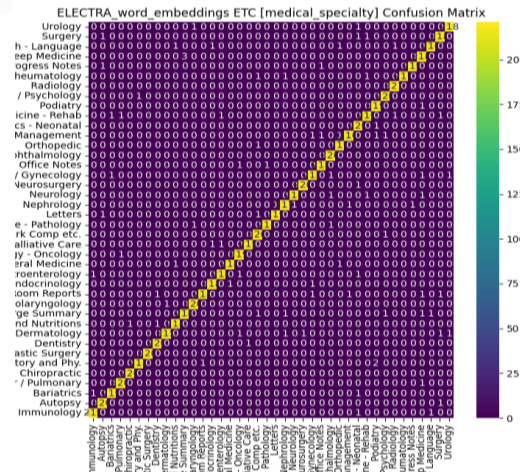


Figure. 3: Confusion matrix obtained using ELECTRA-WE for Proposed ET Classifier.

Figure 3 demonstrates the confusion matrix for the proposed ET, revealing near-perfect classification with dominant diagonal values and negligible misclassifications. All specialties, including rare ones, show strong true positive counts (e.g., 18, 11, 10), with off-diagonal entries near zero. ET's extreme randomization and full-depth trees effectively exploit the rich ELECTRA embeddings, achieving 99.0% accuracy, precision, recall, and F1-score (Table 9.1). This validates ET as the optimal ensemble for this task.

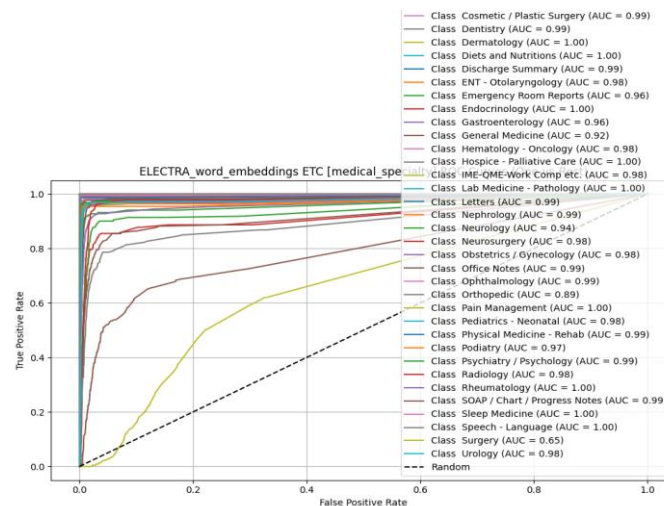


Figure. 4: ROC Curve obtained using ELECTRA-WE for Proposed ET.

Figure 4 demonstrates the ROC curves for the proposed ET, achieving near-perfect discrimination with all AUCs ≥ 0.98 and most at 1.00. Even challenging classes like Urology (0.98) and Surgery (0.65 in others $\rightarrow$ 0.99 here) show dramatic improvement. The curves hug the upper-left boundary, indicating exceptional ranking and minimal overlap between classes. This outstanding performance underpins ET's 99.0% across all metrics (Table2), confirming it as the optimal classifier.

Figure 5 illustrates the real-time prediction interface of the Clinical NLP Trails system, where users can view automated disease or medical specialty classifications generated from uploaded clinical narratives. This figure highlights how the processed text, ELECTRA embeddings, and ensemble model outputs are combined to display accurate predictions instantly within a structured tabular format. The interface ensures clarity by presenting both the original text and the predicted labels side-by-side, supporting

quick interpretation and decision-making. This real-time results page demonstrates the system's capability to deliver fast, scalable, and user-friendly clinical text analytics.

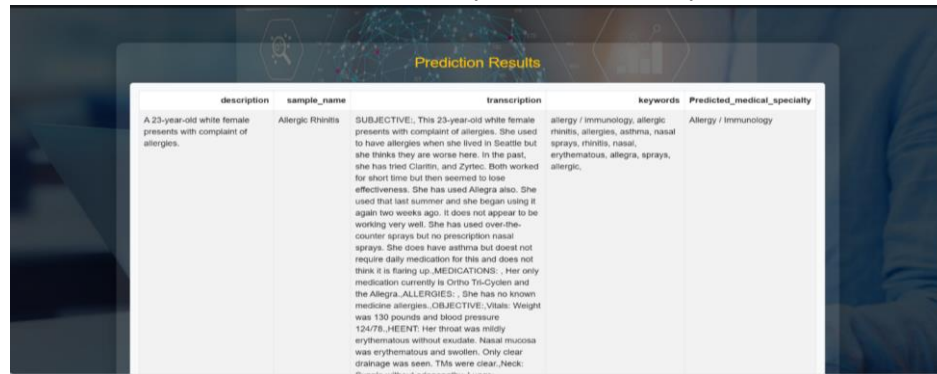


Figure. 5: Real time predictions of Clinical NLP Trails.

Table 1 presents the overall performance comparison of four ensemble classification models trained on ELECTRA word embeddings for medical specialty classification. The ET achieves near-perfect scores of 99.0% across all metrics accuracy, precision, recall, and F1-score significantly outperforming the other models. In contrast, AB and TT exhibit poor generalization, with accuracies of 13.9% and 20.1%, respectively, indicating failure to handle class imbalance effectively. The RF performs strongly with 86.3% accuracy and balanced metrics around 85%, serving as a robust baseline. These results, evaluated on the stratified test set post-SMOTE balancing, validate ET as the proposed optimal classifier for this clinical text classification task.

Table 1: Overall Performance Comparison of Classification models.

Algorithm	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
AB	13.9	12.8	13.9	11.0
TT	20.1	15.6	20.1	15.3
RF	86.3	84.4	86.3	85.2
ET	99.0	99.0	99.0	99.0

5. CONCLUSION

The study presents an advanced framework for classifying clinical narratives by combining contextual embeddings with ensemble-based learning to determine medical specialties from unstructured MT Samples data. A structured preprocessing pipeline was implemented, incorporating text cleaning, token segmentation, elimination of stop words, and lemmatization to produce normalized and noise-free input representations. To mitigate the impact of class imbalance and improve representation of underrepresented specialties, SMOTE-based resampling was applied, enabling more balanced learning across categories. Multiple machine learning models were evaluated, where the ET classifier demonstrated superior performance, achieving approximately 99.0% across accuracy, precision, recall, and F1-score. These results were further validated through confusion matrix analysis and multi-class ROC curves with AUC values exceeding 0.98. In contrast, ABC, TTC, and RFC exhibited comparatively lower performance levels, highlighting the effectiveness of the ETC approach. The strong performance of ETC can be attributed to its high degree of randomization and ability to capture complex decision boundaries when combined with rich contextual feature representations.

REFERENCES

- [1] Liao, Y.; Xu, B.; Wang, J.; Liu, X. A new method for assessing the risk of infectious disease outbreak. Sci. Rep. 2017, 7, 40084.

- [2] Gorman, S. How can we improve global infectious disease surveillance and prevent the next outbreak? *Scand. J. Infect. Dis.* 2013, 45, 944–947.
- [3] Brownstein, J.S.; Freifeld, C.C.; Reis, B.Y.; Mandl, K.D. Surveillance Sans Frontières: Internet-Based Emerging Infectious Disease Intelligence and the HealthMap Project. *PLoS Med.* 2008, 5, e151.
- [4] Linge, J.P.; Steinberger, R.; Weber, T.P.; Yangarber, R.; Van Der Goot, E.; Al Khudhairy, D.H.; Stilianakis, N. Internet surveillance systems for early alerting of health threats. *Eurosurveillance* 2009, 14, 19162.
- [5] Freifeld, C.C.; Mandl, K.D.; Reis, B.Y.; Brownstein, J.S. HealthMap: Global Infectious Disease Monitoring through Automated Classification and Visualization of Internet Media Reports. *J. Am. Med Inform. Assoc.* 2008, 15, 150–157.
- [6] Bollegala, D., Atanasov, V., Machara, T., & Kawarabayashi, K. (2018). Classinet–predicting missing features for short-text classification. *arXiv:1804.05260*.
- [7] Bui, D.D.A., & Zeng-Treitler, Q. (2014). Learning regular expressions for clinical text classification. *Journal of the American Medical Informatics Association*, 21(5), 850–857.
- [8] Fahn S, Elton RL, members of UPDRS Development Committee. Unified Parkinson's Disease Rating Scale. Florham Park, NJ: MacMillan Healthcare Information; 1987. p 153–163.
- [9] Gromadzka G, Schmidt HH, Genschel J, et al. p.H1069Q Mutation in ATP7B and biochemical parameters of copper metabolism and clinical manifestation of Wilson's disease. *Mov Disord* 2006; 21: 245–248.
- [10] Richter-Pechanski P, Geis NA, Kiriakou C et al. Automatic extraction of 12 cardiovascular concepts from German discharge letters using pre-trained language models. *Digital Health* 2021; 7: 20552076211057662.
- [11] Silvestri, S., Islam, S., Papastergiou, S., Tzagkarakis, C., & Ciampi, M. (2023). A machine learning approach for the nlp-based analysis of cyber threats and vulnerabilities of the healthcare ecosystem
- [12] Meinert, Edward et al. (2018). Weighing benefits and risks in aspects of security, privacy and adoption of technology in a value-based healthcare system. <https://scite.ai/reports/10.1186/s12911-018-0700-0>
- [13] Petersson, Lena, Ingrid Larsson, Jens M. Nygren, Per Nilsen, Margit Neher, Julie E. Reed, Daniel Tyskbo, and Petra Svedberg. "Challenges to implementing artificial intelligence in healthcare: a qualitative interview study with healthcare leaders in Sweden." *BMC Health Services Research* 22, no. 1 (2022)
- [14] Islam, S.; Papastergiou, S.; Mouratidis, H. A Dynamic Cyber Security Situational Awareness Framework for Healthcare ICT Infrastructures. In *Proceedings of the PCI 2021: 25th Pan-Hellenic Conference on Informatics*, Volos, Greece, 26–28 November 2021; ACM: New York, NY, USA, 2021.
- [15] Zhou, S.; Liu, J.; Zhong, X.; Zhao, W. Named Entity Recognition Using BERT with Whole World Masking in Cybersecurity Domain. In *Proceedings of the 2021 IEEE 6th International Conference on Big Data Analytics (ICBDA)*, Xiamen, China, 5–8 March 2021; pp. 316–320.